

USING RNGSTREAMS FOR PARALLEL RANDOM NUMBER GENERATION IN C++ AND R

ANDREW T. KARL

Adsurgo LLC, Denver, CO

RANDY EUBANK AND JELENA MILOVANOVIC AND MARK REISER AND
DENNIS YOUNG

Arizona State University, Tempe, AZ

ABSTRACT. The `RngStreams` software package provides one viable solution to the problem of creating independent random number streams for simulations in parallel processing environments. Techniques are presented for effectively using `RngStreams` with C++ programs that are parallelized via `OpenMP` or `MPI`. Ways to access the backbone generator from `RngStreams` in R through the `parallel` and `rstream` packages are also described. The ideas in the paper are illustrated with both a simple running example and a Monte Carlo integration application.

NOTICE

This is a preprint of an article appearing in *Computational Statistics* (in press). The final publication is available at <http://dx.doi.org/10.1007/s00180-014-0492-3>

1. INTRODUCTION

Simulation studies are in many respects the ideal application for parallel processing. They are “naturally parallel” in the sense that computations may generally be conducted without the need for inter-processor communication or modification of the base algorithm. Thus, near linear speedup can be realized in many cases that occur in practice. However, in order for a division of labor across multiple processors to be productive, the random number streams being used by each processor must behave as if they are independent in a probabilistic sense. Fortunately, methods now exist that, when used properly, assure the requisite stream independence. We explore one such approach in this paper that is available through the `RngStreams` software package.

The effect of stream dependence is easy to demonstrate as seen in the output from an R session exhibited below.

```
> x <- NULL
> for (i in 1:200) {
+   set.seed(123)
```

```

+   .Random.seed[3:626] <- as.integer(.Random.seed[3:626] +
+                                     (i - 1) * 100)
+   zTemp <- rnorm(1000)
+   x <- c(x, sqrt(1000) * mean(zTemp))
+ }
> set.seed(123)
> y <- NULL
> for (i in 1:200) {
+   zTemp <- rnorm(1000)
+   y <- c(y, sqrt(1000) * mean(zTemp))
+ }
> cat(shapiro.test(x)$p.value, shapiro.test(y)$p.value, "\n")
0.0001438936 0.4466189

```

Following along the lines of Hechenleitner and Entacher (2003), we have generated sets of 1000 random numbers from the standard normal distribution using 200 different random number streams deriving from R's default Mersenne Twister (MT) random uniform generator. The streams are constructed by using seeds (vectors of length 624) that are spaced 100 units apart in each component. The mean from each stream is then rescaled to produce numbers that have ostensibly been sampled from a standard normal distribution. However, the Shapiro-Wilk test firmly rejects the normality hypothesis. In contrast, data from a similar process that permits the normal evolution of the MT states easily pass the Shapiro-Wilk evaluation. Although the relationship between streams will likely be much more complex in situations that arise in practice, the message remains the same: inter-stream dependence can sabotage a parallel random number generation scheme.

Hill (2010) reviews available methods for parallel random number generation, including random spacing (workers are initialized to randomly spaced positions on the period of the same generator by assigning different, randomly generated seeds), sequence splitting (dividing a sequence into non overlapping contiguous blocks), cycle division (a more involved technique for sequence splitting where the period of a generator is deterministically divided into segments) and parametrization (parameters associated with a generator are varied to produce different streams). The random spacing method carries a risk of overlap between random streams on different processors with an unlucky sampling of seeds, although the risk is small when using generators with large periods.

MT provides one example of a long-period generator with the period $2^{19937} - 1$ (Matsumoto and Nishimura, 1998). This feature along with its widespread availability has made MT a popular choice for parallel generation via random spacing. In spite of this, there are reasons that make the development and use of other parallel generation methods worthwhile. For example, it is generally advisable to repeat simulations with different generators to ensure that the results are not an artifact of the structure of a particular generator (Gentle, 2003). So, there is value in having other available generator options for this purpose. In addition, a method such as cycle division comes with a guarantee that the streams will not collide, whereas random spacing only renders such events unlikely. Lemieux (2009), for example, advocates against random spacing and recommends using generators that are guaranteed not to overlap. As a case in point, she directs us to the work of Matsumoto et al. (2007) who show that many modern random number generators

exhibit correlations if their initial states are chosen using another linear generator with a similar modulus.

Issues that arise when using linear congruential generators for parallel random number generation have been well documented. Random spacing and sequence splitting may result in random number streams that exhibit undesirable dependence properties (e.g., Cenacchi and De Matteis (1970), De Matteis and Pagnutti (1988), Anderson and Titterton (1970), and Etacher (1999)). Similar results have been quantified for some nonlinear generators. For example, De Matteis and Pagnutti (1990) show that long range correlation is present in random number streams produced by sequence splitting with any generator of the form $x_n = f(x_{n-1}) \bmod 2^m$ if the generator is such that the period halves when the modulus halves.

Many of the problems with congruential generators can be avoided by using combined multiple recursive generators (CMRG) with cycle division. A MRG is roughly equivalent to a multiple recursive generator with a period that can be as large as the product of the periods of the generators used in the combination (L'Ecuyer and Tezuka, 1991). When combined appropriately (e.g., as in (3) below), the lattice structure inherent to multiple recursive generators is effectively blurred (L'Ecuyer, 1996). With a good choice for the parameters (e.g., L'Ecuyer (1999)), these combined generators have also fared well when subjected to statistical tests as in L'Ecuyer and Simard (2007). As a case in point, the MRG32k3a generator that provides the backbone generator for the `RngStreams` package developed by L'Ecuyer et al. (2001) is known to have good theoretical and statistical properties. It is cited by Lemieux (2009), along with MT, as a generator that can be “safely used”.

Of course, cycle division becomes an option only if it can be done efficiently. For multiple recursive generators L'Ecuyer (1990) has shown that there are computable matrices that can be used to advance the state of the generator to any specified point in its associated random number stream. This feature is what makes the combination of generators easy to use in a cycle division paradigm. The `RngStreams` package gives an implementation of this approach in the context of its MRG32k3a generator. Haramoto et al. (2008) have recently developed a jump ahead method for MT. The number generation step is two to three times faster for each stream than with `RngStreams`; but, the sequence splitting itself is roughly 1000 times slower.

Another noteworthy package that provides functionality for parallel random number generation is the `SPRNG` package of Mascagni and Srinivasan (2000). In contrast to `RngStreams`, it produces different random number sequences for the processes via parametrization of a single type of generator. For example, one of the package's featured generators is a multiplicative lagged Fibonacci generator whose states fall into 2^{1007} different equivalence classes of generators (each of period 2^{81}) that provide the different streams.

In this paper we focus on the use of the `RngStreams` package. There are several reasons for this. For example, L'Ecuyer et al. (2001) in reference to the `SPRNG` package state that it is “not supported by the same theoretical analysis for the quality and independence of the different streams” as `RngStreams`. There are also practical difficulties that arise in using `SPRNG` due to its ties to MPI and a rather complex implementation that necessitates the creation and linking of a compiled library. In

contrast, the **RngStreams** package is quite compact with all its source code available from <http://www.iro.umontreal.ca/~lecuyer/myftp/streams00/>.

The C++ implementation of **RngStreams** provides an object oriented framework where objects are created which have an associated method **RandU01** that produces uniform random deviates. The overall result is that **RngStreams** is easy to incorporate into C/C++ programs using **OpenMP** or **MPI** through simple include directives. This property extends its applications from cluster down to desktop environments.

General treatments of parallel computing in R are provided by Schmidberger et al. (2009) and Eugster et al. (2011). In contrast to these papers, the R segments of this article are directed toward the specific task of parallel random number generation in the language. In this regard, the **RngStreams** generators become available in R through the **rstream** package (Leydold, 2012). They are also the default generator for the **parallel** package that provides multithreading capabilities and now comes as part of the R base software.

The goal of this article is to illustrate how easily **RngStreams** can be employed in parallel settings in C++ with **OpenMP** and **MPI** and in R (R Core Team, 2013) through the **rstream** and **parallel** packages. Our initial presentation involves only the most elementary features of these APIs. However, it is our belief that these simple programs contain fundamental techniques that can provide a foundation for safe parallel random number generation in much more complex code that might evolve in practice. An application of Monte Carlo integration in the context of an item trait model is used to support this contention.

In the next section we describe the basic features of the **RngStreams** generator. Then, in subsequent sections we address its use in C++ and R. These sections are connected through a common example that allows us to illustrate both the differences and similarities in using **RngStreams** under the different APIs.

2. RNGSTREAMS

The “backbone” generator for **RngStreams**, often referred to as MRG32k3a, is a combination of the two multiple recursive generators that produce the states

$$\begin{aligned} (1) \quad x_{1,n} &= (1403580 \times x_{1,n-2} - 810728 \times x_{1,n-3}) \bmod (4294967087), \\ (2) \quad x_{2,n} &= (527612 \times x_{2,n-1} - 1370589 \times x_{2,n-3}) \bmod (4294944443) \end{aligned}$$

at the n th step of the recursion given initial seeds $\tilde{x}_{i,0} = (x_{i,-2}, x_{i,-1}, x_{i,0})^T, i = 1, 2$. The two states are combined to produce the uniform random deviate u_n via the rule

$$\begin{aligned} (3) \quad z_n &= (x_{1,n} - x_{2,n}) \bmod 4294967087, \\ u_n &= \begin{cases} z_n/4294967088, & \text{if } z_n > 0, \\ 4294967087/4294967088, & \text{if } z_n = 0. \end{cases} \end{aligned}$$

This particular generator is capable of producing a random number stream of period length (roughly) 2^{191} under the condition that $x_{1,0}, x_{1,-1}, x_{1,-2}$ are not all 0 and are all less than 4294967087 and $x_{2,0}, x_{2,-1}, x_{2,-2}$ are all less than 4294944443 and not all 0. For use in a parallel context this long cycle is divided into 2^{64} non-overlapping streams each of length 2^{127} .

The key to accessing different streams produced by (1)-(2) is the technique described in L’Ecuyer (1990) that allows movement between streams through linear transformations of the initial seeds using known matrices. Specifically, if $\tilde{x}_{1,0}, \tilde{x}_{2,0}$

are the initial seeds/states for the two generators, at the n th step of the recursion the state for generator i is $\tilde{x}_{i,n} = (A_i^n \bmod m_i) \tilde{x}_{i,0} \bmod m_i, i = 1, 2$, for known 3×3 matrices A_1, A_2 with $m_1 = 4294967087, m_2 = 4294944443$. The powers of the matrices A_1, A_2 can be computed explicitly and, in particular the matrices A_1^{127}, A_2^{127} that are needed for movement between the different streams of the generator have been calculated and are contained in an anonymous namespace in the `RngStream.cpp` file. This feature is used in the MPI section of our paper.

The initial state of the package is set using the function `SetPackageSeed` that has prototype

```
static bool SetPackageSeed (const unsigned long seed[6])
```

with `seed` containing the six initial seeds for the generator that must be supplied by the user. The first `RngStream` object that is created will use this initial seed. Subsequent objects will then be started on the next stream; i.e., they will have initial “seeds” that have been advanced 2^{127} states from the initial state of the previous object’s generator. Our aim in the sequel is to perform some simple experiments that will illustrate how the seeds are changed for the different processes that are working concurrently in a parallel computing environment.

3. OPENMP

OpenMP represents the standard API for shared memory parallel processing. An overview of the various OpenMP functions is provided, e.g., in Chapman et al. (2001). Here we will use only the API features that allow us to parallelize a block of code.

The listing below illustrates the use of `RngStreams` in an OpenMP program.

LISTING 1. `ranOpenMP.cpp`

```
//ranOpenMP.cpp
#include <omp.h>
#include "RngStream.h"
#include <iostream>

int main(){

    int nP = omp_get_num_procs();
    omp_set_num_threads(nP); //set number of threads

    unsigned long seed[6] = {1806547166, 3311292359,
                             643431772, 1162448557,
                             3335719306, 4161054083};
    RngStream::SetPackageSeed (seed);
    RngStream RngArray[nP]; //array of RngStream objects

    int myRank;
    #pragma omp parallel private(myRank)
    {
        myRank = omp_get_thread_num();
        #pragma omp critical
        {
            std::cout << RngArray[myRank].RandU01() << " ";
        }
    }
    std::cout << std::endl;
    return 0;
}
```

The first step is to include the `OpenMP` header file `omp.h` via an include directive. The number of processors `nP` is determined and used to set the number of threads (with `omp_get_num_procs` and `omp_set_num_threads`, respectively). The `SetPackageSeed` function is then used to set the seeds for the `RngStreams` package. This particular choice of seeds was made to provide results that can be compared to those obtained from `R` in Section 5.

The desired random deviates are produced with an array of `RngStream` objects: one object for each thread. The `RngStream` class default constructor is designed so that the first object will have the seed designated with `SetPackageSeed` while subsequent objects will have initial states/seeds have been advanced 2^{127} states from that of their predecessor.

An alternative approach that one might try here is to simply create a unique `RngStream` object for each thread with a `private` specification. In such instances the default constructor will also be called. However, we have experienced cases where a race condition can lead to threads with duplicate seeds. The array approach ensures that all objects are created correctly while effectively exploiting the shared memory feature that allows all threads access to the array elements.

A random deviate is generated from each of the `RngStream` objects using the class method `RandU01` that invokes the backbone generator. These values will serve as subsequent comparison benchmarks. The actual generation of random numbers in the parallel region is quite simple with each thread using the object in `RngArray` that corresponds to its unique thread number assigned by the `OpenMP` API. The resulting values are printed out from within a `critical` region to avoid garbled output. We compile and execute Listing 1 using the gnu compiler suite on a four processor machine in the following manner.

```
$ g++ -fopenmp ranOpenMP.cpp RngStream.cpp -o ranOpenMP
$ ./ranOpenMP
0.341106 0.312399 0.166374 0.149433
```

4. MPI

The **Message Passing Interface** or `MPI` provides a number of functions that can be used for inter-processor communication; see, e.g., Quinn (2003). Our development here will use only the simple communication paradigm where a master and worker processes transmit and receive messages via the `MPI send` and `recv` functions.

As noted at the beginning of this section, the seeds/states of `RngStream` objects are advanced by multiplication involving known matrices that are available as part of the package. A function that can be used to carry out these multiplications is `matVecModM` that has prototype

```
void MatVecModM(const double A[3][3], const double s[3],
               double v[3], double m)
```

This function will perform multiplication of a 3×3 array `A` times a three-element array `s` modulo `m` and return the result in the three-element array `v`. For our purposes we will take `m` to be either $m_1 = 4294967087$ or $m_2 = 429494443$, `A` to be either A_1^{127} or A_2^{127} and `s` to be a seed vector that was allocated to a previous process. The moduli and matrices (as well as `MatVecModM`) are contained in an

anonymous namespace within `RngStream.cpp` under the names `m1`, `m2`, `A1p127` and `A2p127`. Thus, the following function will manually advance the input array `seedIn` 2^{127} states to `seedOut` which occupies the same relative location in the next stream.

```
static void AdvanceSeed(unsigned long seedIn[6],
                       unsigned long seedOut[6]){
    double tempIn[6]; double tempOut[6];
    for(int i = 0; i < 6; i++) tempIn[i] = seedIn[i];
    MatVecModM (A1p127, tempIn, tempOut, m1);
    MatVecModM (A2p127, &tempIn[3], &tempOut[3], m2);
    for(int i = 0; i < 6; i++)
        seedOut[i] = tempOut[i];
}
```

We wrapped this function with the necessary supporting code from `RngStream.cpp` in a supplemental class `RngStreamSupp` whose header file `RngStreamSupp.h` appears in the code below.

Our MPI implementation of the `RngStreams` methodology now takes the following form.

LISTING 2. `ranMPI.cpp`

```
//ranMPI.cpp
#include <iostream>
#include "mpi.h"
#include "RngStream.h"
#include "RngStreamSupp.h"

int main(){
    unsigned long seed[6] = {1806547166, 3311292359,
                             643431772, 1162448557,
                             3335719306, 4161054083};

    MPI::Init();//start MPI
    int myRank = MPI::COMM_WORLD.Get_rank();//get process rank
    int nP = MPI::COMM_WORLD.Get_size();//get number of processes
    if(myRank == 0){
        //generate deviate for master process
        RngStream::SetPackageSeed(seed);
        RngStream Rng;
        double U = Rng.RandU01();
        //now send seeds to the other processes
        unsigned long tempSeed[6];
        for(int i = 1; i < nP; i++){
            RngStreamSupp::AdvanceSeed(seed, tempSeed);
            MPI::COMM_WORLD.Send(tempSeed, 6, MPI::UNSIGNED_LONG,
                                i, 0);
            for(int j = 0; j < 6; j++)
                seed[j] = tempSeed[j];
        }
        std::cout << U << " ";
        //collect the other random deviates
        for(int i = 1; i < nP; i++){
            MPI::COMM_WORLD.Recv(&U, 1, MPI::DOUBLE, i, 0);
            std::cout << U << " ";
        }
        std::cout << std::endl;
    }
    else{
```

```

    unsigned long mySeed[6];
    //receive your seed if you are not the master process
    MPI::COMM_WORLD.Recv(mySeed, 6, MPI::UNSIGNED_LONG, 0, 0);
    RngStream::SetPackageSeed(mySeed);
    RngStream Rng;
    double U = Rng.RandU01();
    //send deviate to master process
    MPI::COMM_WORLD.Send(&U, 1, MPI::DOUBLE, 0, 0);
}
MPI::Finalize(); //end MPI
return 0;
}

```

The idea behind Listing 2 is straightforward. The master process uses an initial seed in conjunction with `AdvanceSeed` to determine seeds that will produce “independent” random number streams for the other processes. These seeds are communicated to the processes using the `send` and `recv` functions. The processes then use them to initialize their particular `RngStream` objects and generate a random uniform value. The communication process is then reversed and the worker processes send their random numbers back to the master for printing.

A typical compile and run sequence for our MPI program might look like

```

$ mpicxx ranMPI.cpp RngStream.cpp RngStreamSupp.cpp -o ranMPI
$ mpiexec -np 4 ranMPI
0.166374 0.341106 0.312399 0.149433

```

This output agrees with our prior results from the `OpenMP` program.

An alternative approach that leads to the same result as in `ranMPI.cpp` is to mimic the array formulation that was used in `ranOpenMP.cpp`. This reduces the amount of communication with the expense of more memory usage for the individual processes. We illustrate this approach in Section 6.

5. RNGSTREAMS IN R

The `RngStreams` backbone generator is now one of the random number generator options in R. For example,

```

> RNGkind("L'Ecuyer-CMRG")
> set.seed(123)
> runif(1)
[1] 0.1663742

```

sets the R uniform random number generator as MRG32k3a, sets the session’s seed and reproduces one of the random deviates from the previous two sections.

At this point it is perhaps worthwhile to mention how we have aligned the seeds that are being used in our R and C++ code. The `set.seed` function in R uses a single integer to set the seed for the current uniform random number generator. The result can then be accessed through the integer vector `.Random.seed`. In terms of our example with the MRG32k3a generator, the vector has seven elements: the first one indicates the type of generator and the remaining six are the seeds. Thus,

```

> set.seed(123)
> cat(.Random.seed[2:7], "\n")
1806547166 -983674937 643431772 1162448557 -959247990 -133913213

```

reveals R’s representation for the six seeds that are required for MRG32k3a. Initially the presence of negative integers seems a bit puzzling; the `RngStreams` seeds are

stored as `unsigned long` in the code for the package. However, in R they are treated as 32-bit unsigned integers internally but otherwise viewed as signed integers with a two's-complement representation. Thus, one must add 2^{32} to any negative integers that appear in `.Random.seed` to see the actual seeds that were used with the generator. For example,

```
> .Random.seed[3] + 2^{32}
[1] 3311292359
```

gives us the second seed that appeared in our C++ code. Although this may appear to be an esoteric detail, it is essential for cross-language output comparison.

Another way to access the MRG32k3a generator is through the `rstream` package that allows us to create instances of a derived class `rstream.mrg32k3a` for the backbone generator. By default, subsequent `rstream.mrg32k3a` objects are initialized to the next stream. In fact, `rstream` will return an error if more than one generator requests a specific seed in the same R session. To override this behavior one must specify the option `force.seed = TRUE`. The following code initializes four generators and uses them with the `rstream.sample` function to each produce a uniform random deviate.

```
> library(rstream)
> rngList <- c(new("rstream.mrg32k3a", seed = c(1806547166,
+       3311292359, 643431772, 1162448557, 3335719306,
+       4161054083), force.seed = TRUE),
+       replicate(3, new("rstream.mrg32k3a")))
> sapply(rngList, rstream.sample)
[1] 0.1663742 0.3411064 0.3123993 0.1494334
```

Note that we have again produced the four independent streams that arose in the previous two sections.

The MRG32k3a generator is a built in fixture of the `parallel` package that offers essentially the same parallel computing functionality as can be obtained separately from the `snow` (Tierney et al., 2013) and `multicore` (Urbanek, 2011) packages. In addition, `parallel` is now bundled within the R base distribution which suggests it will be one of the key ingredients in the evolution of R's future parallel processing capabilities.

The function `makeCluster` from the `parallel` package will create a (default, socket) cluster whose seed can then be set with `clusterSetRNGStream`. The code below illustrates this process using a cluster of size four.

```
> library(parallel)
> cl <- makeCluster(4)
> clusterSetRNGStream(cl, 123)
> parSapply(cl, rep(1, 4), runif)
[1] 0.1663742 0.3411064 0.3123993 0.1494334
> stopCluster(cl)
```

The random numbers were generated using `parSapply` that gives a parallel version of `sapply`; there are parallel versions of `lapply` and `apply` as well as several other related options. Upon completion of our task, the cluster is terminated with `stopCluster`. The output confirms that we are using the MRG32k3a generator and that the seeds are being advanced correctly for each node in the cluster.

The other way that `parallel` provides multithreading ability (for non-Windows machines) is through `mclapply`, which furnishes a parallelized version of `lapply`

suited for a shared memory environment. An attempt to replicate our cluster results using this approach produces

```
> library(parallel)
> RNGkind("L'Ecuyer-CMRG")
> set.seed(123)
> unlist(mclapply(rep(1,3),runif))
[1] 0.3411064 0.3123993 0.9712727
> runif(1)
[1] 0.1663742
```

which illustrates that the seeds for new streams are not being advanced as expected. Note that the call `mclapply(rep(1,3),runif)` produces three calls to `runif` with `n=1` and collects the results in a list. This issue is easy enough to fix with

LISTING 3. `mclapply` with built-in L'Ecuyer CMRG

```
> library(parallel)
> RNGkind("L'Ecuyer-CMRG")
> set.seed(123)
> unlist(mclapply(rep(1, 3), runif, mc.cores = 3))
[1] 0.3411064 0.3123993 0.1494334
> runif(1)
[1] 0.1663742
```

It is important to specify the desired number of additional cores (beyond the master process) via the `mc.cores` option. This value is not automatically set to the length of the first argument of `mclapply`. On Windows platforms, `mc.cores` must be set to 1 since `mclapply` relies on the Unix fork command, which is not available in Windows.

It is possible to gain more direct control over the behavior of the random streams on each core using the `rstream` package. This provides a valuable option when reproducibility across multiple languages and platforms is desired. For example,

LISTING 4. `mclapply` with `rstream`

```
> library(rstream)
> library(parallel)
> rngList <- c(new("rstream.mrg32k3a",
+               seed = c(1806547166, 3311292359, 643431772,
+               1162448557, 3335719306, 4161054083),
+               force.seed = TRUE),
+             replicate(3, new("rstream.mrg32k3a")))
> unlist(mclapply(rngList, rstream.sample, mc.cores = 4))
[1] 0.1663742 0.3411064 0.3123993 0.1494334
```

returns the results we had anticipated. On Windows platforms, where `mc.cores` must be 1, the program in Listing 3 will produce numbers from a single stream from `RngStreams`. By contrast, the code from Listing 4 will faithfully reproduce the results from a multicore non-Windows platform during serial execution on a Windows machine. A similar approach of explicitly creating the streams with `rstream` may also be used with the cluster approach in order to maximize transparency.

6. AN APPLICATION

In this section the basic ideas in Sections 3–5 are applied to a real problem involving Monte Carlo integration. The resulting code listings now become rather lengthy with the consequence that we can only discuss a few key aspects here.

The setting is that of Bock and Lieberman (1970) where a latent trait model (LTM) is being fit to data obtained from subject responses on the LSAT. There are five dichotomous variables $Y = (Y_1, \dots, Y_5)$ whose observed values $y = (y_1, \dots, y_5)$ produce $2^5 = 32$ possible response patterns. If y is one such pattern, it is assumed that the conditional probability of seeing y given the value z of a latent variable $Z \sim N(0, 1)$ takes the form

$$(4) \quad p(y|z) = \prod_{i=1}^5 p_i(z)^{y_i} (1 - p_i(z))^{1-y_i},$$

where

$$(5) \quad p_i(z) = 1 / (1 + \exp \{ \alpha_i + \beta z \})$$

with $\alpha = (\alpha_1, \dots, \alpha_5)$ a vector of variable specific intercepts and β a common slope parameter. This leads to

$$(6) \quad \begin{aligned} p(y) &:= E_z[p(y|z)] \\ &= \int_{-\infty}^{\infty} p(y|z) \phi(z) dz \end{aligned}$$

with ϕ the standard normal density. We will focus on approximation of this integral via Monte Carlo methods. Other integrals such as those obtained by differentiation of (6) with respect to α and β that arise in the computation of marginal maximum likelihood estimators can be handled analogously.

The choice of α, β is arbitrary for our purposes and, accordingly, we take them to be the marginal likelihood estimates returned from the `grm` function in the `R ltm` package (Rizopoulos, 2006); it is these values that will be seen in our code. What requires somewhat more justification is our focus on an example with a single latent variable. In general, there may be many latent variables in an LTM that make the integral in (6) multidimensional. As detailed in Section 1.1 of Lemieux (2009), the number of points in product-rule quadrature must grow exponentially with the dimension of the integrand to maintain a constant rate of decay for the approximation error. In contrast, the expected error from a Monte Carlo integral estimator is independent of dimension and is always of order $1/\sqrt{n}$ with n the number of sampling points. Thus, one concludes that Monte Carlo integration is more appropriate in higher dimensions and, for example, univariate integrals are better and more precisely evaluated by deterministic quadrature. On the other hand, the basic premise that underlies Monte Carlo integration remains the same regardless of the dimension: repeatedly generate values from a probability distribution, evaluate the integrand at these values and average the results in some general sense. Consequently, there is little conceptual downside to our use of a single latent variable in this instance while the payoff in terms of code length and mathematical simplification is rather substantial.

For any specified pattern y , a Monte Carlo estimator of (6) will typically have the form

$$(7) \quad \text{probFunc}(y) = \sum_{i=1}^n W(z_i) p(y|z_i)$$

for z_1, \dots, z_n standard normal deviates and W a suitable weight function. The serial aspect of our computation problem is primarily about evaluation of (7). We will therefore deal with that issue first before turning to the parallelization step. In both cases the R and C++ code will be developed in tandem to better illustrate the similarities and differences that arise when attempting to solve this type of problem in the two different languages.

The C++ code (included in `funcClass.cpp` in the supplementary material) that evaluates (7) is

LISTING 5. `probFunc` in C++

```
TNT::Array1D<double>
funClass::probFunc(const TNT::Array1D<double>& Z,
                  const TNT::Array2D<int>& patn){
    //pattern probability array used in intermediate calculations
    TNT::Array2D<double> probMat(patn.dim1(), patn.dim2(), 0.);
    //work array
    TNT::Array1D<double> temp(patn.dim1(), 0.);
    //array that will hold the cell probabilities
    TNT::Array1D<double> prob(patn.dim1(), 0.);

    //sum across normal deviates
    for(int k = 0; k < Z.dim(); k++){
        probMat = probMatFunc(Z[k], patn);

        //multiply to get the probability for each pattern
        for(int i = 0; i < patn.dim1(); i++){
            temp[i] = 1.;
            for(int j = 0; j < patn.dim2(); j++){
                temp[i] *= probMat[i][j];
            }
            temp[i] *= 2.*dnorm(Z[k])
                /(dnorm(Z[k]/2.)*(double) Z.dim());
        }
        prob += temp;
    }
    return prob;
}
```

This function takes and returns arguments from the Template Numerical Toolkit (TNT) that provides implementations of vector (as `Array1D` objects) and matrix (as `Array2D` objects) classes with overloaded addition, multiplication, etc., operators. These array classes are implemented as templates which has the consequence that all code is placed in header files that can be downloaded directly from <http://math.nist.gov/tnt>.

The call to `probMatFunc` that appears here corresponds to a function that evaluates $p_i(z)^{y_i}(1 - p_i(z))^{1-y_i}$ across all (32 in this instance) possible response patterns for a given value of z and returns the result as an array with columns representing the different variables in the model. The values that are used for the z argument are passed into the function in the `Array1D` object `Z`. These are chosen to be pseudo-random normal deviates with 0 mean and standard deviation 2. This feature in conjunction with the weights that are applied at the end of the most interior (or `i`) loop produces a simple importance sampler estimate of the target integral. The code for `probFunc`, `probMatFunc` and several utility functions (e.g., the standard normal density and quantile function designated as `dnorm` and `qnorm`, respectively,

in our code) have been collected into a C++ class `funClass` that contains a class member to hold the values of the coefficients.

The R version of Listing 5 (see `LTM_parallel.R` in the supplement) is embodied in three functions. First, there is

```
> PFunc <- function(z, coefVec, patn) {
+   ncells <- nrow(patn)
+   probMat <- ProbMatFunc(z, coefVec, patn)
+   p <- matrix(0, ncells, 1)
+   # multiply the probabilities across variables
+   p <- as.matrix(dnorm(z) * apply(probMat, 1, prod), ncells, 1)
+   p
+ }
```

that evaluates (4) across all response patterns (i.e., y values) for a given value of the latent variable (i.e., z). The call to `ProbMatFunc` in this instance is for the R analog of its C++ namesake in Listing 5. The `patn` argument is an input array that holds the response patterns as in the C++ case. However, in contrast to the C++ development, in this instance the coefficient vector that is needed for evaluating (5) is treated as an input variable, named `coefVec`, that is passed to this and other functions that use it directly from a driver program.

Next, the function `PFunc` is vectorized via

```
> VecPFunc <- Vectorize(PFunc, vectorize.args = "z")
```

to handle the array case and thereby allow us to use it with `apply` for “averaging” across normal deviates as in

LISTING 6. `ProbFunc` in R

```
> ProbFuncparLapply <- function(nSim, nP, coefVec, patn) {
+   Z <- qnorm(runif(round(nSim/nP, 0)), 0, 2)
+   prob <- colMeans(apply(VecPFunc(Z, coefVec = coefVec,
+                                   patn = patn),
+                           1, function(y) y/dnorm(Z, 0, 2)))
+ }
```

The `Z` array that appears in this listing is, again, a vector of pseudo-random normal deviates with 0 mean and standard deviation 2 that is used in the same weighting scheme as for the C++ code.

The C++ MPI code (available from `driverLTM_mpi.cpp` in the supplement) that calculates (7) in parallel looks like

```
RngStream::SetPackageSeed(seed);
RngStream RngArray[nP];

//array for random numbers
TNT::Array1D<double> Z(nSim[myRank], 0.);
//funClass object
funClass f(pAlpha, beta, patn.dim2());

//array that will hold approximate probabilities
TNT::Array1D<double> prob(patn.dim1(), 0.);

//approximate the cell probabilities on each process
for(int i = 0; i < nSim[myRank]; i++)
    Z[i] = 2.*f.qnorm(RngArray[myRank].RandU01());
```

```

prob = f.probFunc(Z, patn);

//now gather up the results on the master process
double* temp;
temp = new double[nP*patn.dim1()];

double* pData_ = new double[patn.dim1()];
for(int i = 0; i < patn.dim1(); i++)
    pData_[i] = prob[i];

MPI::COMM_WORLD.Gather(pData_, patn.dim1(), MPI::DOUBLE,
                       temp, patn.dim1(), MPI::DOUBLE, 0);

if(myRank == 0){
    std::cout << "Integration results:"<<std::endl;
    for(int i = 0; i < prob.dim(); i++){
        for(int j = 1; j < nP; j++){
            prob[i] += temp[j*prob.dim() + i];
        }
        prob[i] /= (double)nP;
        std::cout<<prob[i]<<std::endl;
    }
    std::cout << "time with " << nP << " processes was " <<
        MPI::Wtime() - wtime << std::endl;
}

```

Here we have employed the scheme from Section 3 for obtaining independent streams: each processes creates the same array of `RngStream` objects and then uses the array element that correspond to its rank. The number of normal deviates that must be generated by the different processes is portioned out in an array `nSim` that is used in the same manner. The `funcClass` object `f` then provides the means for each process to transform the uniform deviates obtained from the `RngStream` object into standard normals and call `probFunc` to approximate the integral.

The use of a common `RngStream` object array has allowed us to circumvent the need for sending information to the processes. It is only necessary to collect up the results they produce. One means to accomplish this directly, without looping, is by use of MPI's `Gather` function which is the path we have chosen in this instance. All processes call this function with a pointer argument for the quantity that is to be passed to the master (or 0) process. This data is then received (by the master) in a pointer of dimension sufficient to hold the collective result (i.e., the product of the number of processes `nP` and the length `prob.dim()` of the `prob` array). The pointer that is passed from the worker processes needs to be one that points to the actual data that is held in the `prob` array. This resides in a pointer named `data_` that is a private member of the `Array1D` class. However, its content is accessible through an overloaded `[]` operator and we used this feature to manually transfer the information in `data_` to the pointer `pData_` that was passed to `Gather`. Once all the integral approximations are collected, the master process averages them across processes to obtain the final approximation.

The `OpenMP` analog of our MPI code is shown in the next listing, and is available from `driverLTM_openmp.cpp` in the supplement.

```

//pointer to Array1D objects that will hold output
//from each process
TNT::Array1D<double>* tempVec = new TNT::Array1D<double>[nP];

```

```

//pointer for random numbers and funClass object
double* pZ;
funClass* pf;
//initialize array for cell probabilities
TNT::Array1D<double> prob(patn.dim1(), 0.);

//approximate the cell probabilities
#pragma omp parallel private(myRank, pZ, pf)
{
    myRank = omp_get_thread_num();
    pZ = new double[nSim[myRank]];
    pf = new funClass(pAlpha, beta, patn.dim2());
    for(int i = 0; i < nSim[myRank]; i++)
        pZ[i] = 2.*pf->qnorm(RngArray[myRank].RandU01());
    tempVec[myRank]
        = pf->probFunc(TNT::Array1D<double>(nSim[myRank],
                                            pZ), patn);

    delete[] pZ;
    delete pf;
}

for(int i = 0; i < nP; i++){
    prob += tempVec[i];
}

std::cout << "Integration results:"<<std::endl;
for(int i = 0; i < patn.dim1(); i++){
    prob[i] /= (double)nP;
    std::cout<<prob[i]<<std::endl;
}

std::cout <<"time with " << nP
    << " processes was " << omp_get_wtime() - wtime
    << std::endl;

```

An array of `RngStream` objects is used to generate the random uniforms for each process exactly as in Section 3. What differs from the MPI treatment is the way we have used pointers for the `funClass` object, the array of standard normals, and storage of the output from `probFunc`. By designating the first two of these pointers as `private` in the `private` clause at the start of the parallel section, we avoid any race conditions that might arise in calling constructors. Each thread can safely create the object and array they will need to perform their task. The now familiar mapping of a processes' rank to an array element index is used to store the `Array1D` objects that each thread produces with `probFunc` in a safe location provided by the `Array1D` pointer. After all threads collapse back to the master at the end of the parallel region, their output is averaged using an overloaded addition assignment operator for the `Array1D` class.

In the R setting our parallel implementation takes the form

```

> library(parallel)
> DriverLTMparLapply <- function(patn, coefVec, nSim, nP) {
+   cl <- makeCluster(nP)
+   clusterSetRNGStream(cl, 123)

```

```

+   clusterExport(cl, c("patn", "coefVec", "nSim", "nP"))
+   clusterExport(cl, c("ProbMatFunc", "PFunc", "VecPFunc",
+   "ProbFuncparLapply"))
+   result <- parLapply(cl, seq_len(nP),
+   function(...) ProbFuncparLapply(nSim,
+   nP, coefVec, patn))
+   stopCluster(cl)
+   # now average across threads
+   prob <- vector(mode = "double", length = nrow(patn))
+   for (i in 1:nP) {
+     prob <- prob + result[[i]]
+   }
+   prob/nP
+ }

```

in a cluster context and

```

> library(parallel)
> library(rstream)
> DriverLTMmclapply <- function(patn, coefVec, nSim, nP) {
+   rngList <- c(new("rstream.mrg32k3a",
+   seed = c(1806547166, 3311292359, 643431772,
+   1162448557, 3335719306, 4161054083),
+   force.seed = TRUE),
+   replicate(nP - 1, new("rstream.mrg32k3a")))
+   result <- mclapply(rngList, ProbFuncmclapply, nSim, nP,
+   coefVec, patn, mc.cores=nP)
+   # now average across processes
+   prob <- vector(mode = "double", length = nrow(patn))
+   for (i in 1:nP) {
+     prob <- prob + result[[i]]
+   }
+   prob/nP
+ }

```

with

```

> ProbFuncmclapply <- function(rngObj, nSim, nP, coefVec, patn) {
+   Z <- qnorm(rstream.sample(rngObj, round(nSim/nP, 0)), 0, 2)
+   prob <- colMeans(apply(VecPFunc(Z, coefVec = coefVec,
+   patn = patn),
+   1, function(y) y/dnorm(Z, 0, 2)))
+ }

```

for shared memory purposes. Note that the various constants and functions must be communicated to the workers in a cluster with `clusterExport` and that we have used the `rstream` package in the same manner as Section 5 to ensure proper stream initialization when using `mclapply`. Both `parLapply` and `mclapply` return lists that hold the output from each of the workers. The list element are then averaged to obtain the final approximation.

The programs print matching output for the integral approximations when the number of threads is held constant. For example, using four threads with the R `mclapply` function produces

```

> DriverLTMmclapply(patn,coefVec,nSim,4)
[1] 0.010591032 0.014465339 0.003321068 0.007459930 0.008497401

```



```
[6] 0.019087239 0.004382200 0.016301170 0.004348307 0.009767360
[11] 0.002242468 0.008341667 0.005737658 0.021343283 0.004900161
[16] 0.031011491 0.012769320 0.028683011 0.006585274 0.024496295
[21] 0.016849315 0.062677083 0.014389904 0.091068923 0.008622165
[26] 0.032073241 0.007363630 0.046601970 0.018840844 0.119237442
[31] 0.027375482 0.310245431.
```

The C++ OpenMP program with four threads produces identical output, as expected:

```
$ g++ -fopenmp RngStream.cpp funclass.cpp RngStreamSupp.cpp
    driverLTM_openmp.cpp -o openmp_ltm
$ ./openmp_ltm 4
Integration results:
0.010591032 0.014465339 0.003321068 0.007459930 0.008497401
0.019087239 0.004382200 0.016301170 0.004348307 0.009767360
0.002242468 0.008341667 0.005737658 0.021343283 0.004900161
0.031011491 0.012769320 0.028683011 0.006585274 0.024496295
0.016849315 0.062677083 0.014389904 0.091068923 0.008622165
0.032073241 0.007363630 0.046601970 0.018840844 0.119237442
0.027375482 0.310245431.
```

Though not shown, the C++ MPI and R `parLapply` programs also produce the same output. When the number of threads are changed, the approximations change as well since the random numbers are generated from different streams.

We carried out a few performance comparisons using the MPI and OpenMP C++ code and the `mcapply` and `parLapply` R code for the Monte Carlo integration application. The average speedups (i.e., the ratios of serial to parallel run times) and run times (in parentheses) are given in Table 1. In all cases we used five runs with 10^5 sampling points (i.e., `nSim` = 10^5) and used 1, 2, 4 and 8 processors. While all of the programs perform well, near-linear speedup is achieved with `mcapply` in R (using the `rstream` package) and in with MPI in C++. As expected, the C++ programs execute at least an order of magnitude faster than the R programs.

TABLE 1. Speedups and (run times, in seconds) for C++ and R code

Processors	mcapply	parLapply	MPI	OpenMP
1	1.000	1.000	1.000	1.000
	(77.941)	(82.541)	(2.254)	(2.199)
2	2.070	1.950	2.014	1.430
	(37.651)	(42.319)	(1.119)	(1.538)
4	4.117	3.597	4.002	3.043
	(18.930)	(22.944)	(0.563)	(0.723)
8	7.884	5.549	7.545	6.595
	(9.886)	(14.874)	(0.299)	(0.333)

As a closing thought, we mention that the accuracy obtained with 10^5 samples can be roughly compared to that of an 18 (uniformly spaced) point trapezoidal quadrature rule in one dimension. However, a grid of 10^{10} points for a product trapezoidal rule would be required to remain competitive with the same Monte Carlo scheme in 8 dimensions, for example.

7. SUMMARY

The `RngStreams` software package provides one freely available solution for creating independent random number streams for simulation experiments that are conducted in a parallel computing environment. The goal of this paper has been to provide an introduction to its use in both distributed and shared memory settings. We have provided minimal working examples along with a more detailed application to illustrate the potential of `RngStreams`.

While `RngStreams` is easy to use, some care is necessary to ensure that streams are distributed correctly to different processes. We have accomplished this by using an array of `RngStream` objects where the array index is employed to uniquely determine which process may access each object. This scheme is guaranteed to produce independent streams in both `OpenMP` and `MPI`. In the latter context, we have described another option that creates seeds on the master process and then distributes them to all the worker processes. There are memory savings from this latter approach that come with the cost of additional inter-processor communication.

In R we demonstrated the use of `RngStreams` through both the `parallel` and `rstream` packages. We explored the built-in functionality for `RngStreams` in `parallel` and demonstrated how `rstream` may be used within `parallel` to improve the portability of the software.

ACKNOWLEDGEMENTS

Some of the computations were carried out on the Saguaro cluster at Arizona State University. The authors are grateful for comments from the anonymous referees that lead to improvements in the paper.

REFERENCES

- Anderson, N. and Titterton, D. (1970), “Cross correlation between simultaneously generated sequences of pseudo-random uniform deviates,” *Statist. Comput.*, 16, 11–15.
- Bock, R. and Lieberman, M. (1970), “Fitting a response model for n dichotomously scored items,” *Psychometrika*, 35, 179–197.
- Cenacchi, G. and De Matteis, A. (1970), “Pseudo-Random Numbers for Comparative Monte Carlo Calculations,” *Numer. Math.*, 16, 11–15.
- Chapman, B., Jost, G., and van der Pas, R. (2001), *Using OpenMP: Portable Shared Memory Parallel Programming*, Cambridge: MIT Press.
- De Matteis, A. and Pagnutti, S. (1988), “Parallelization of Random Number Generators and Long-Range Correlation,” *Numer. Math.*, 53, 595–608.
- (1990), “Long-Range Correlation in Linear and Non-Linear Random Number Generators,” *Parallel Comput.*, 14, 207–210.
- Etacher, K. (1999), “On the CRAY-system random number generator,” *Simulation*, 72, 163–169.
- Eugster, M. J. A., Knaus, J., Porzelius, C., Schmidberger, M., and Vicedo, E. (2011), “Hands-on Tutorial for Parallel Computing with R,” *Comput Stat*, 26, 219–239.
- Gentle, J. (2003), *Random Number Generation and Monte Carlo Methods*, New York: Springer.

- Haramoto, H., Matsumoto, M., Nishimura, T., Panneton, F., and L'Ecuyer, P. (2008), "Efficient jump ahead for \mathcal{F}_2 -linear random number generators," *INFORMS J Comput*, 20, 385–390.
- Hechenleitner, B. and Entacher, C. (2003), "Pitfalls When Using Parallel Streams in OMNeT++ Simulations," in *Proceedings of Inter-domain Performance and Simulation Workshop, Salzburg, Austria, 11-20*.
- Hill, D. R. C. (2010), "Practical Distribution of Random Streams for Stochastic High Performance Computing," in *International Conference on High Performance Computing and Simulation (HPCS)*, pp. 1–8.
- L'Ecuyer, P. (1990), "Random Numbers for Simulation," *Comm ACM*, 33, 85–98.
- (1996), "Combined multiple recursive random number generators," *Oper Res*, 44, 816–822.
- (1999), "Good parameters and implementation for combined multiple recursive random number generators," *Oper Res*, 47, 159–164.
- L'Ecuyer, P. and Simard, R. (2007), "TestU01: A C Library for Empirical Testing of Random Number Generators," *ACM Trans. Math. Softw.*, 33, 22.
- L'Ecuyer, P., Simard, R., Chen, J., and Kelton, W. (2001), "An Object-Oriented Random-Number Package With Many Long Streams and Substreams," Tech. rep., University of Montreal.
- L'Ecuyer, P. and Tezuka, S. (1991), "Structural properties for two classes of random number generators," *Math Comp*, 57, 735–746.
- Lemieux, C. (2009), *Monte Carlo and Quasi-Monte Carlo Sampling*, New York: Springer.
- Leydold, J. (2012), *rstream: Streams of Random Numbers*, R package version 1.3.2.
- Mascagni, M. and Srinivasan, A. (2000), "Algorithm 806: SPRNG: A Scalable Library for Pseudorandom Number Generation," *ACM Trans Math Software*, 26, 436–461.
- Matsumoto, M. and Nishimura, T. (1998), "Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator," *ACM Transactions on Modeling and Computer Simulation*, 8, 3–30.
- Matsumoto, M., Wada, I., Kuramoto, A., and Ashihara, H. (2007), "Common Defects in Initialization of Pseudorandom Number Generators," *ACM Transactions on Modeling and Computer Simulation*, 17(4):15.
- Quinn, M. (2003), *Parallel Programming in C with MPI and OpenMP*, New York: McGraw Hill.
- R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Rizopoulos, D. (2006), "ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses," *Journal of Statistical Software*, 17, 1–25.
- Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L., and Mansmann, U. (2009), "State of the Art in Parallel Computing with R," *J Stat Software*, 31, 1.
- Tierney, L., Rossini, A. J., Li, N., and Sevcikova, H. (2013), *snow: Simple Network of Workstations*, R package version 0.3-12.
- Urbanek, S. (2011), *multicore: Parallel processing of R code on machines with multiple cores or CPUs*, R package version 0.1-7.